

**UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK**

IN RE:

OPENAI, INC.,
COPYRIGHT INFRINGEMENT LITIGATION

This Document Relates To:

All Actions

25-md-3143 (SHS) (OTW)

Hon. Sidney H. Stein
Hon. Ona T. Wang

NOTICE OF OPENAI'S TECHNOLOGY TUTORIAL DEMONSTRATIVES



OpenAI's Technology Tutorial

In re OpenAI Copyright MDL
MDL No. 25-md-3143 (S.D.N.Y.)

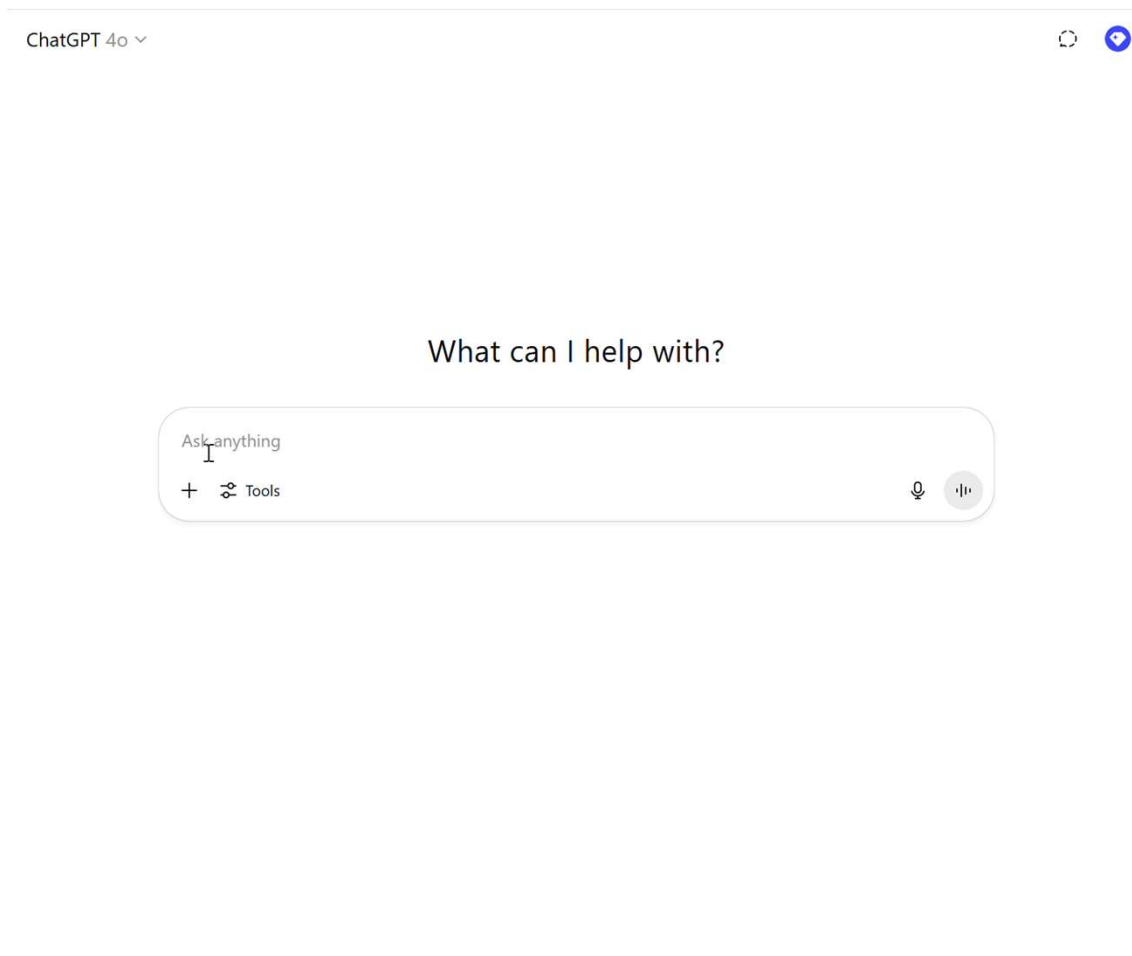
June 26, 2025

What is a GPT?

GENERATIVE

PRE-TRAINED

TRANSFORMER



ChatGPT **generates** innovation



NASA-GPT: Searching the Entire NASA Technical Reports Server Using AI

SUPERCOMPUTING

“The future applications within NASA are nearly endless.”



What is the size of the inlet bleed holes on the XB-70?

ChatGPT uses Large Language Models (LLMs): “GPTs”

A GPT is a statistical model that predicts the next word by learning patterns in language

Prompt:

How do humans learn language?

Response:

Through patterns, observation, and social

Interactions 23%

Feedback 11%

Engagement 10%

Media 6%

What is a GPT?

GENERATIVE

— *Not a database; generates **new** content*

PRE-TRAINED

— *Trained to identify linguistic patterns*

TRANSFORMER

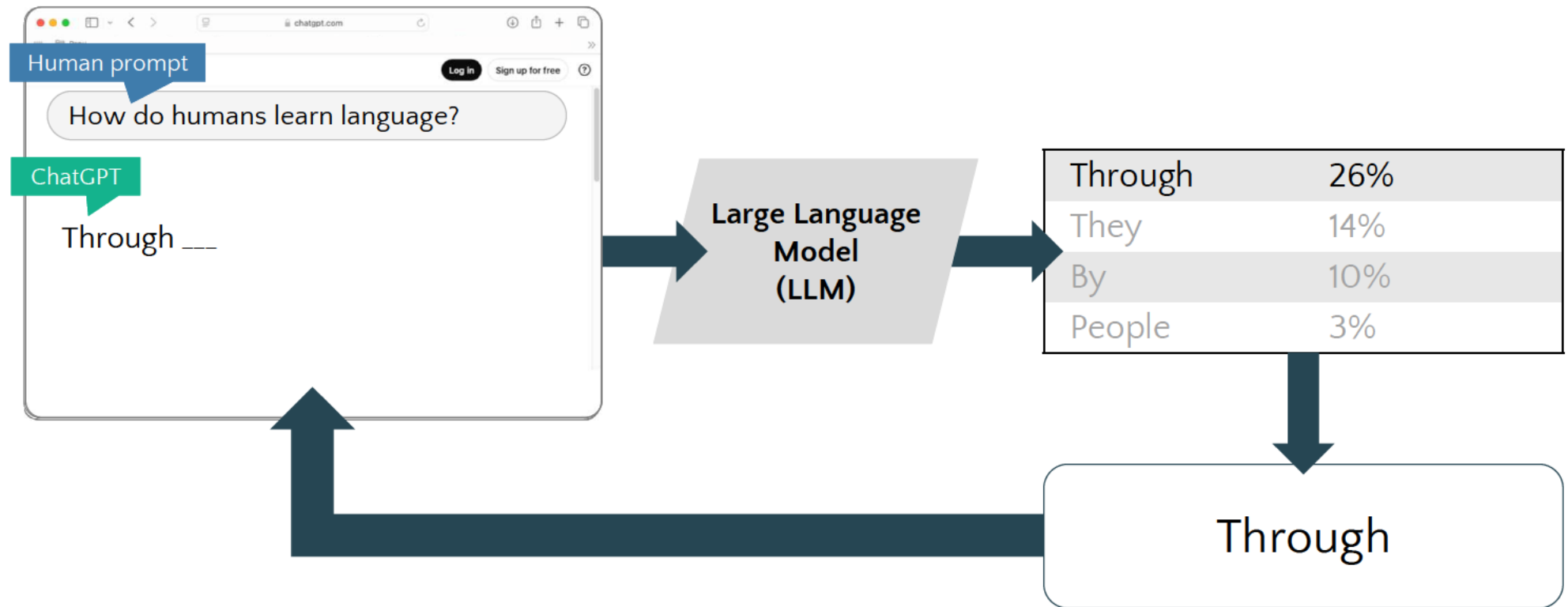
— *Transforms language data into a platform for **innovation***

LLMs are not databases or libraries

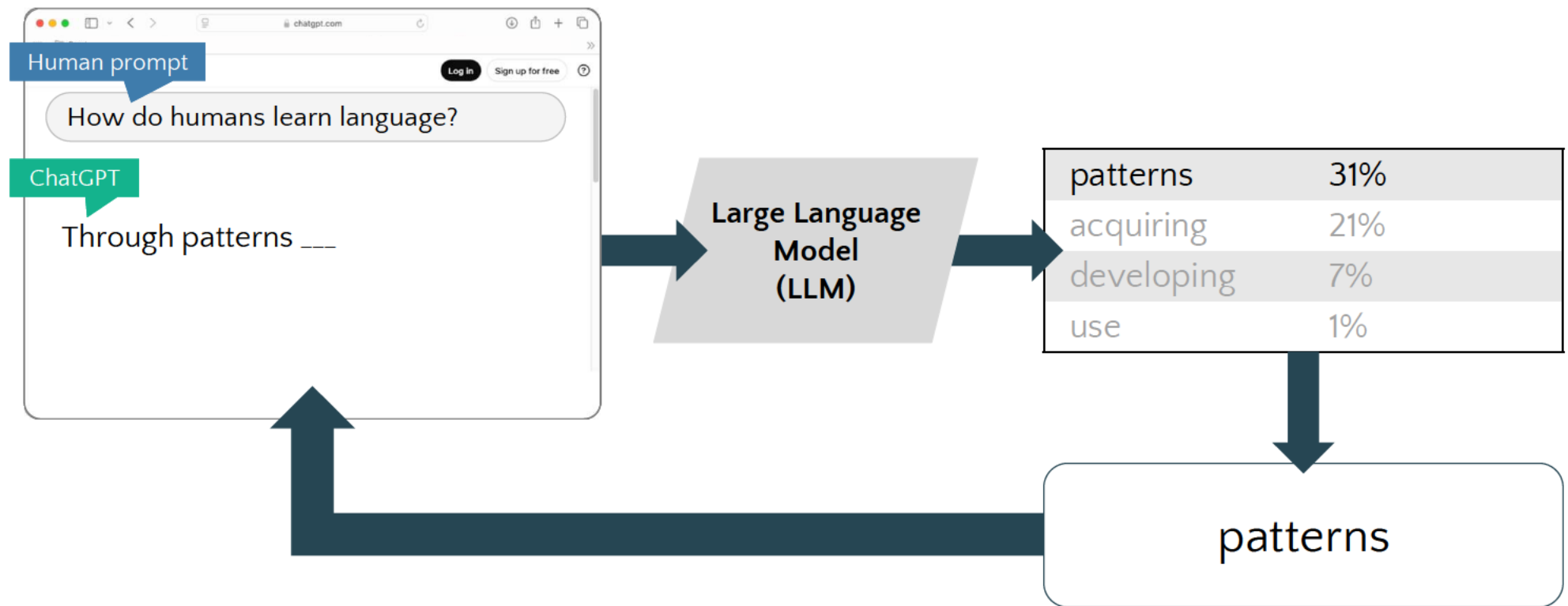
They do not store information or retrieve information from a database



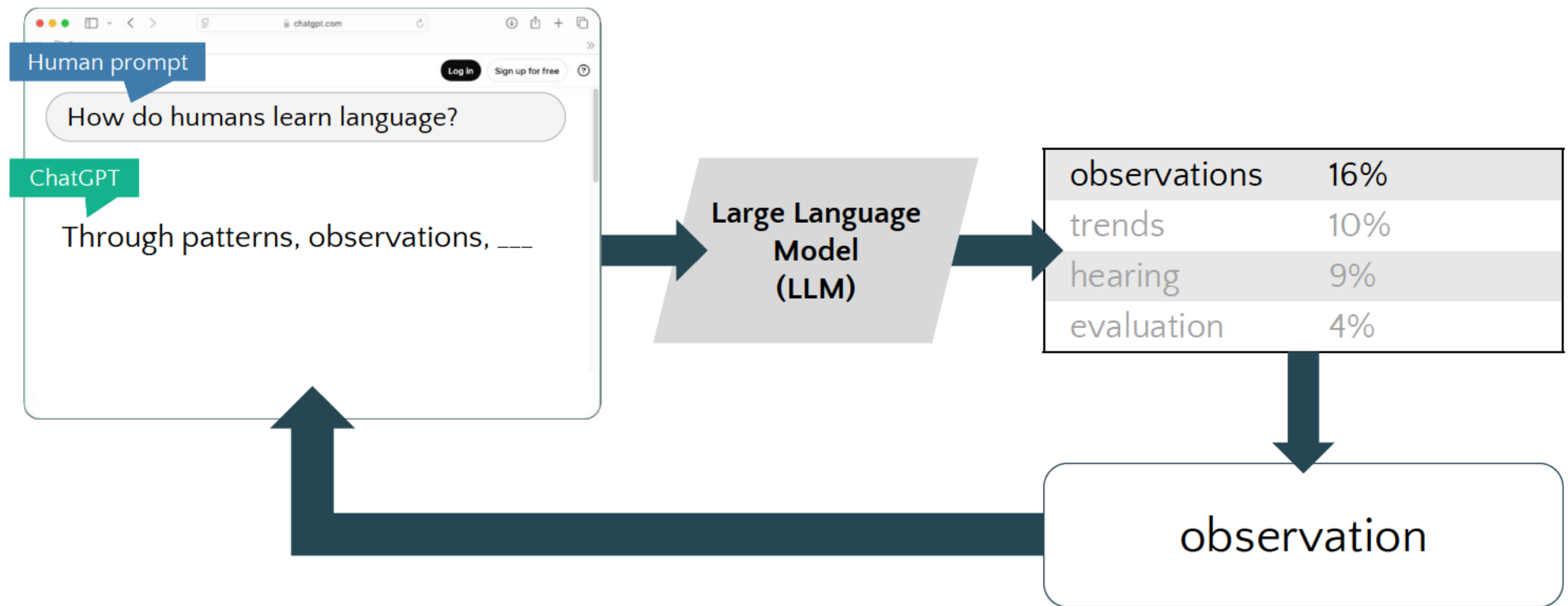
LLMs output probabilities of what words might come next



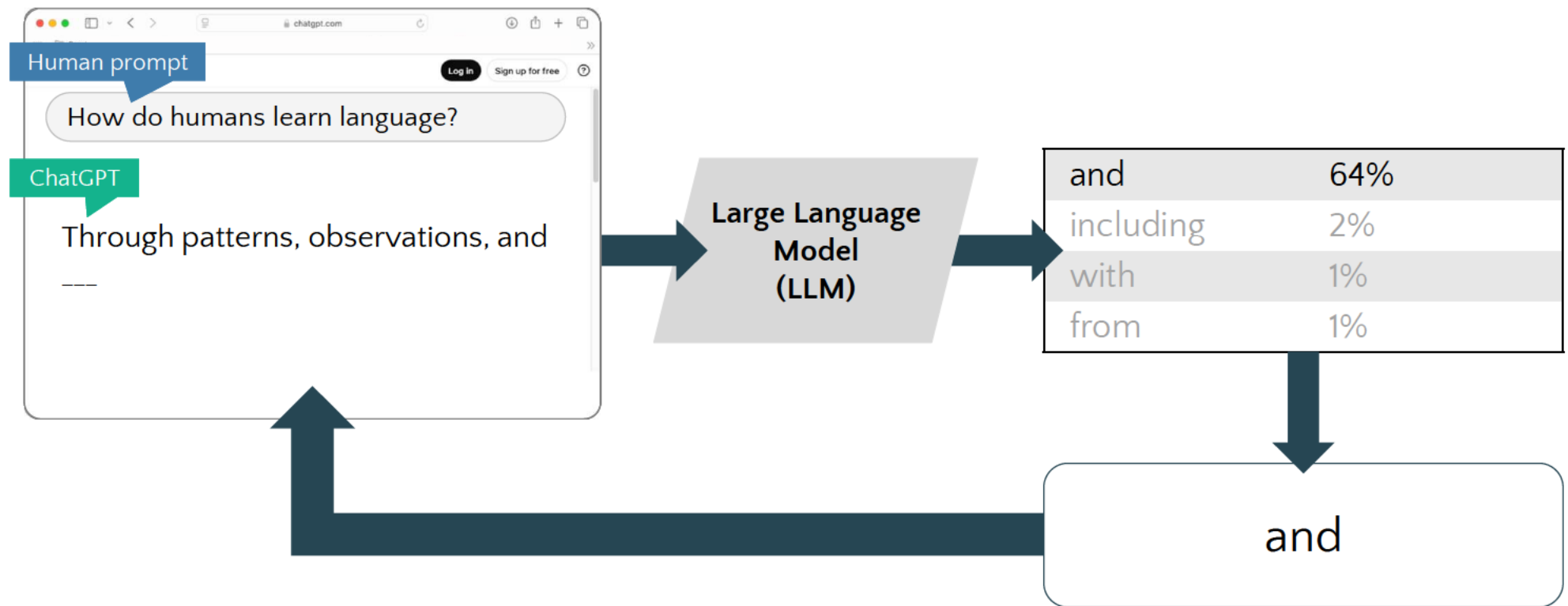
LLMs output probabilities of what words might come next



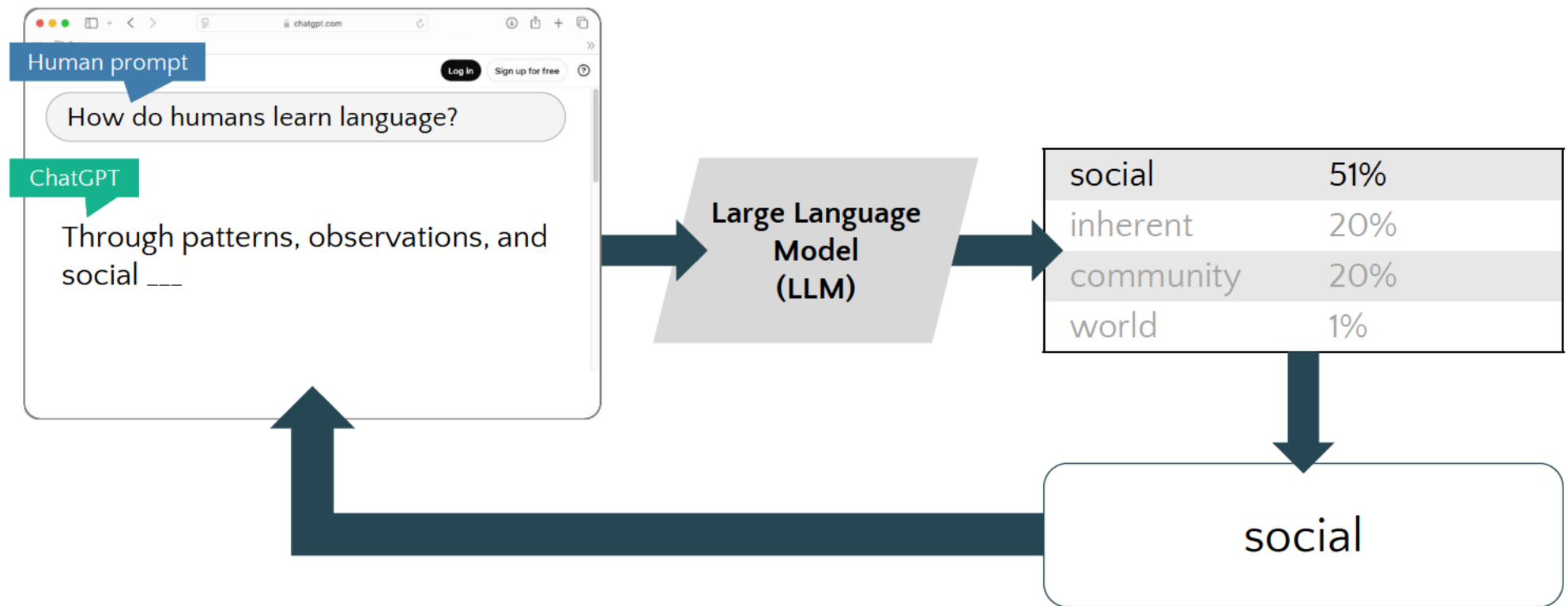
LLMs output probabilities of what words might come next



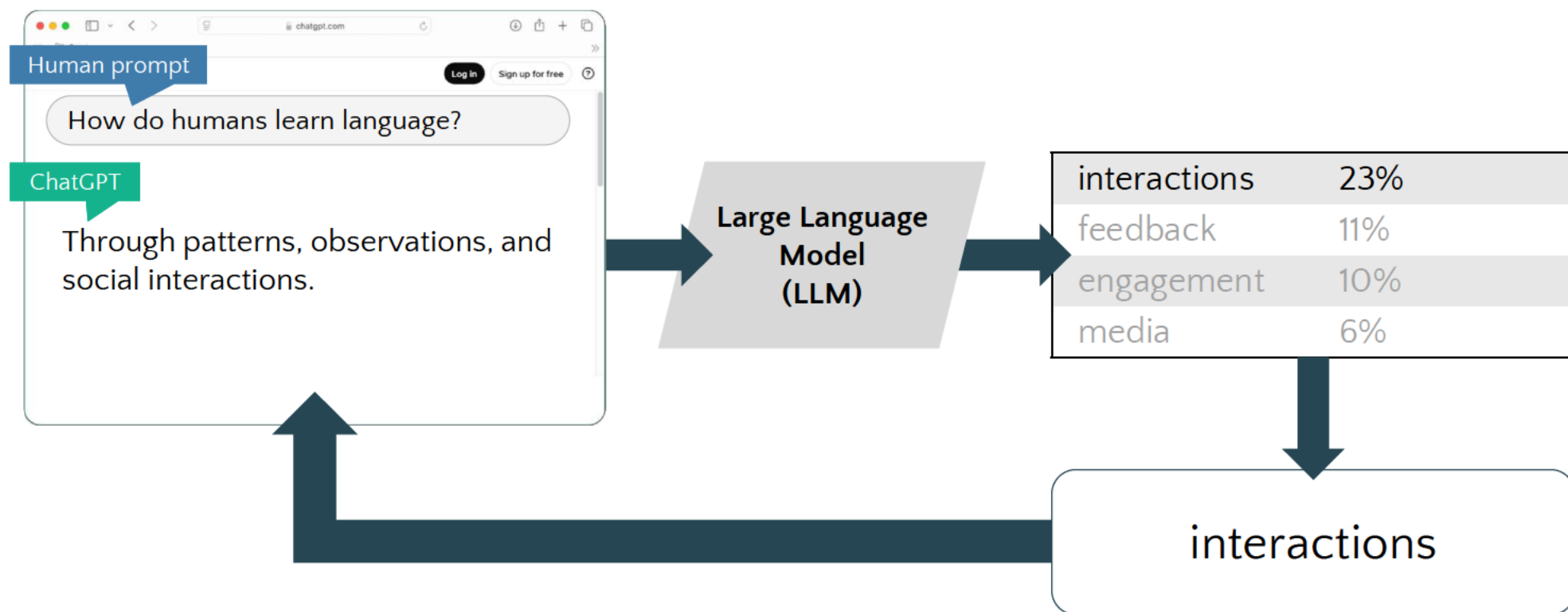
LLMs output probabilities of what words might come next



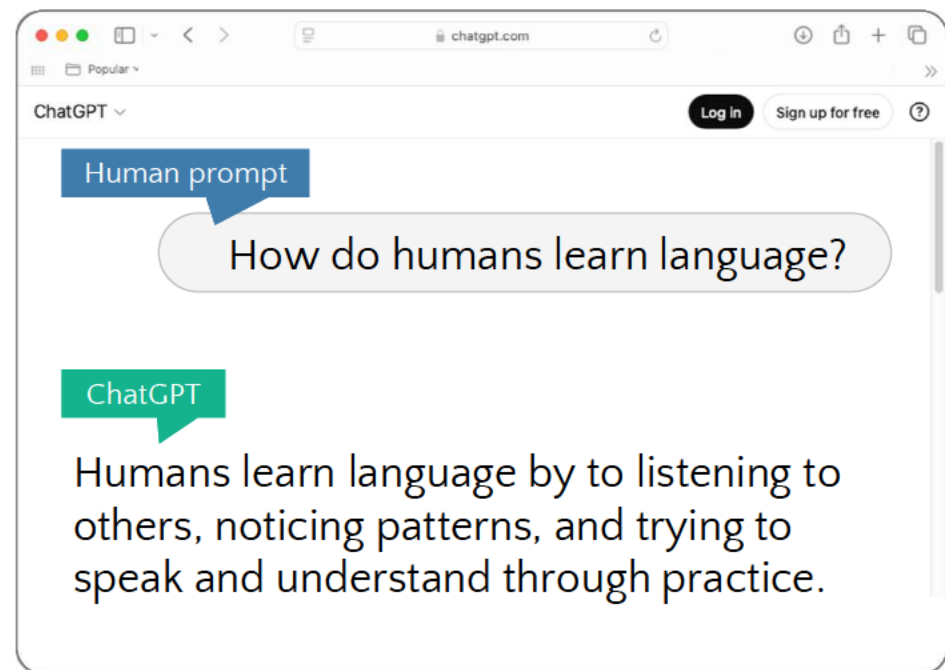
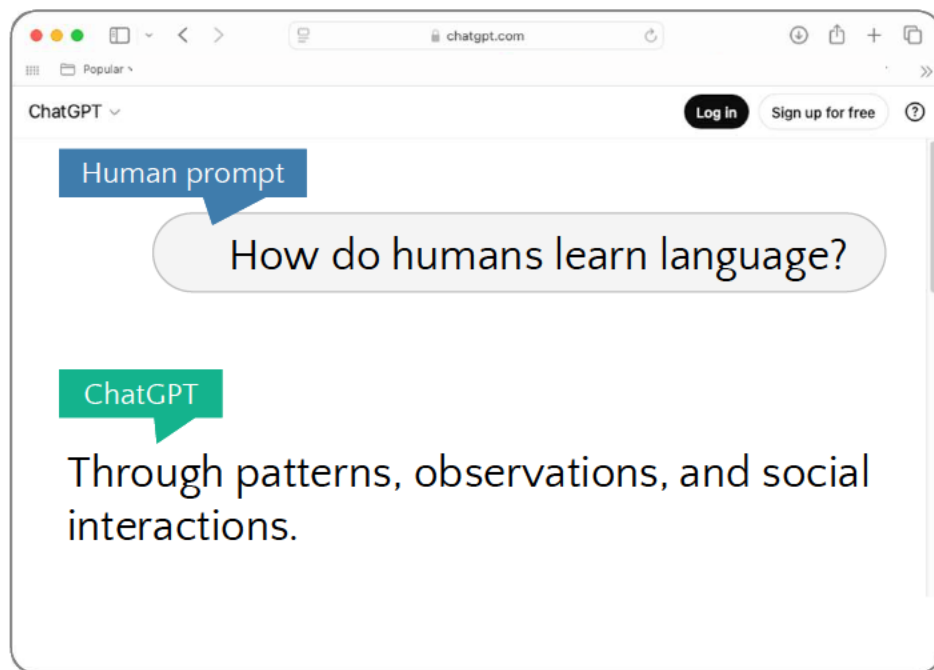
LLMs output probabilities of what words might come next



LLMs output probabilities of what words might come next



Because LLMs are not designed to memorize, the same input can result in different outputs



What is a GPT?

GENERATIVE

— *Not a database; generates **new** content*

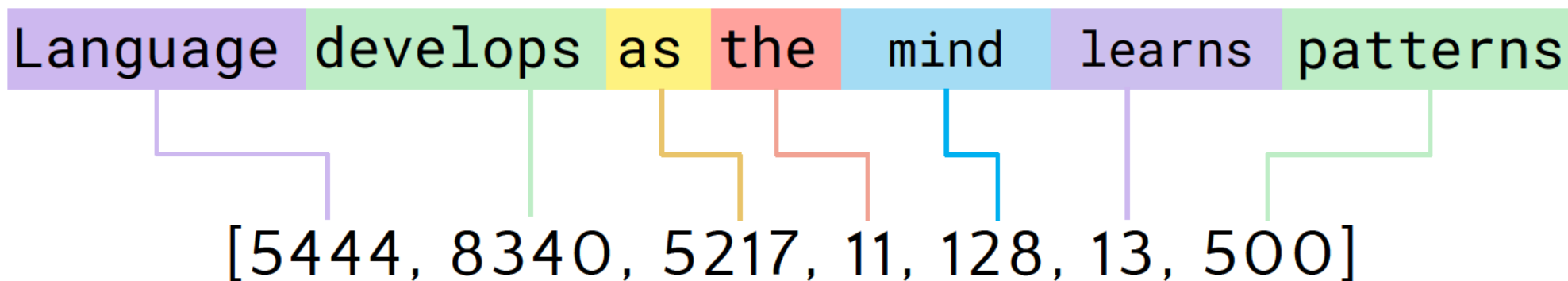
PRE-TRAINED

— ***Trained** to identify linguistic **patterns***

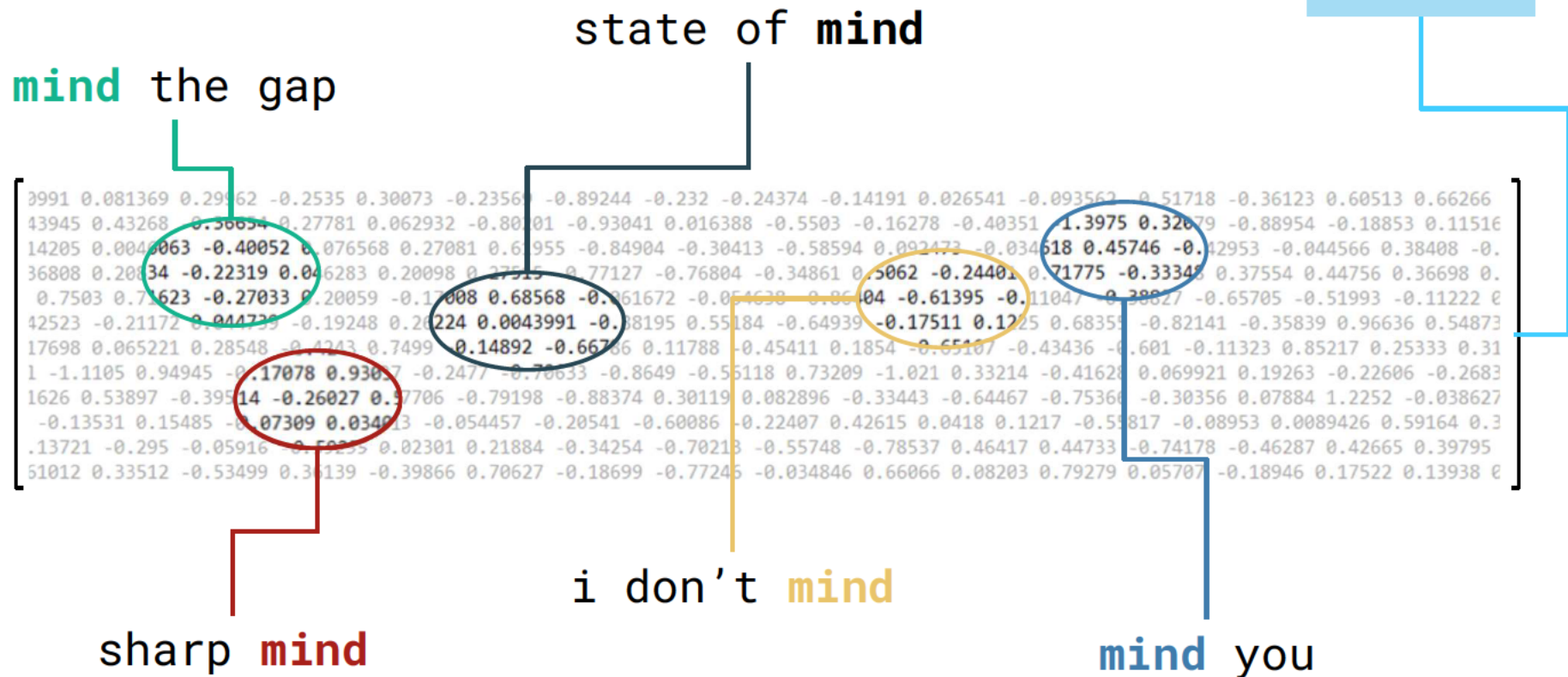
TRANSFORMER

— ***Transforms** language data into a platform for **innovation***

During **tokenization**, words are transformed into numbers that a computer can understand



Each word = thousands of numbers



Each word = thousands of numbers

brain

```
[ -0.65099 1.26331 -0.87651 -1.72957 1.97517 -0.15320 0.11393 -0.84078 0.63132 1.76751 -1.89129 0.74010 0.11836 -0.67188 -1.77762
7 0.58466 0.42617 0.31083 -1.45158 -0.61226 -1.20194 0.27056 1.19842 -1.71352 -1.15130 0.78276 -1.98115 0.24613 0.77700 -0.35849
1.43668 -0.64877 0.40269 -1.26063 -1.25985 -0.23417 -0.58839 0.57958 0.85748 0.20365 -1.83983 -1.79238 -0.01656 1.11073 -1.8417.
1.44323 1.31257 0.54199 0.29543 -1.70525 1.79376 0.91679 -0.15764 1.35187 -0.10371 1.57899 0.28532 0.84626 -1.44230 0.02376 -1..
-0.83522 -1.83013 0.07620 1.02852 0.59177 -1.29791 -1.89422 1.87762 -0.52827 -1.05487 1.98063 -1.53682 -1.27855 -0.80357 0.03492
0.59323 0.97371 -0.15144 -0.87792 0.01189 -0.47492 1.62576 0.89438 -0.62645 1.87121 -0.14262 -1.62334 -0.50791 -1.62259 -1.67862
0.55613 1.06068 -0.16104 0.59806 0.11356 -0.02945 1.05590 -1.27190 1.86503 -1.34049 0.01791 -1.63649 1.75626 1.60784 -1.68794 -1.68794
-1.53942 1.10862 0.40227 -1.81691 0.33745 -0.91474 -0.25783 -0.72632 -1.39547 -1.65267 0.69399 0.57808 -1.15688 1.01552
1.15709 -1.39353 -0.17078 0.93009 0.39702 -1.66281 1.75238 -0.12592 0.65880 -0.93610 -0.49126 -1.35556 -0.68012 -0.93368 -1.86732
-1.31560 1.77706 14 -0.26027 0.3999 -0.17338 0.42629 -0.02297 0.73985 -0.91590 0.59878 1.63638 -1.72895 -0.25977 -1.25119 -1.63072
-0.28878 -0.61643 -0.07309 0.03409 1.7769 -1.51774 -1.15856 1.21638 0.68062 1.15233 -1.03632 1.07117 -1.77285 -0.54255 -0.94361 -1.86732
0.07620 1.02852 0.59177 -1.29791 -1.89422 1.87762 -0.52827 -1.05487 1.98063 -1.53682 -1.27855 -0.80357 0.03492 0.94425 0.59323
-1.45158 -0.61226 -1.20194 0.27056 1.19842 -1.71352 -1.15130 0.78276 -1.98115 0.24613 0.77700 -0.35849 -1.00467 1.43668 -0.64877
```

quick **brain**

The model learns statistical relationships between tokens

- The model is shown billions of sequences of tokens and asked:

“Given this sequence, what token is most likely to come next?”

- The model runs these tests over and over, billions of times.

Training is about adjusting the weights, not about memorizing training data

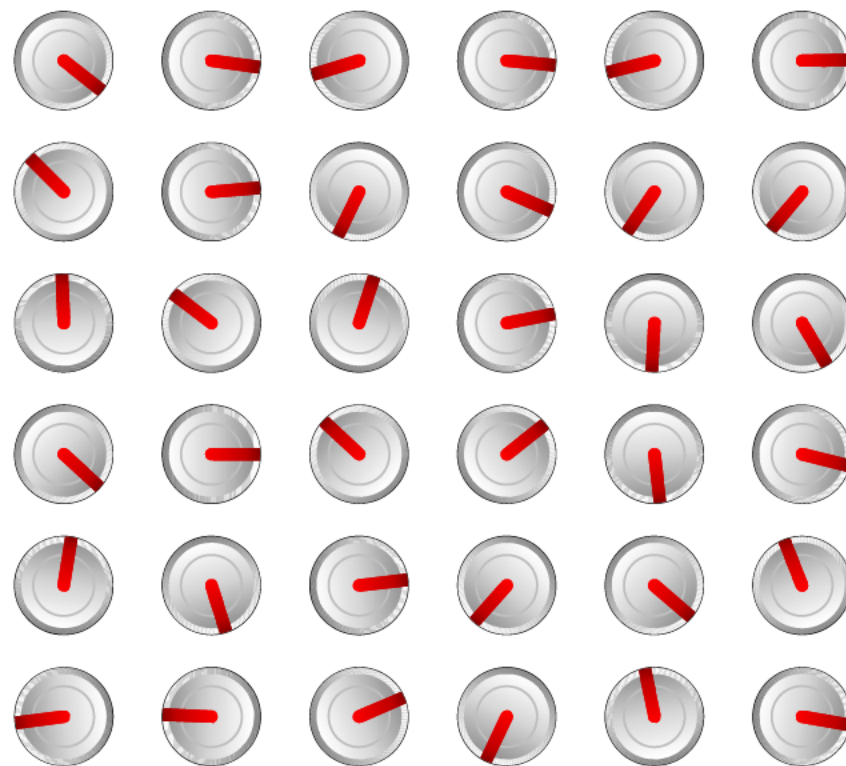
“...patterns emerge from repeated phrases...”

“...pathways strengthen with every utterance...”

“...babies observe linguistic patterns...”

“...babbling refines phoneme perception...”

Training Data



Weights

Goal: teach the model to **infer** concepts not seen before

“...most **birds** lay their **eggs** in carefully constructed nests...”

“...a **birds** will lay three to five **eggs** per clutch...”

“...unlike mammals, **birds** reproduce by laying **eggs**...”

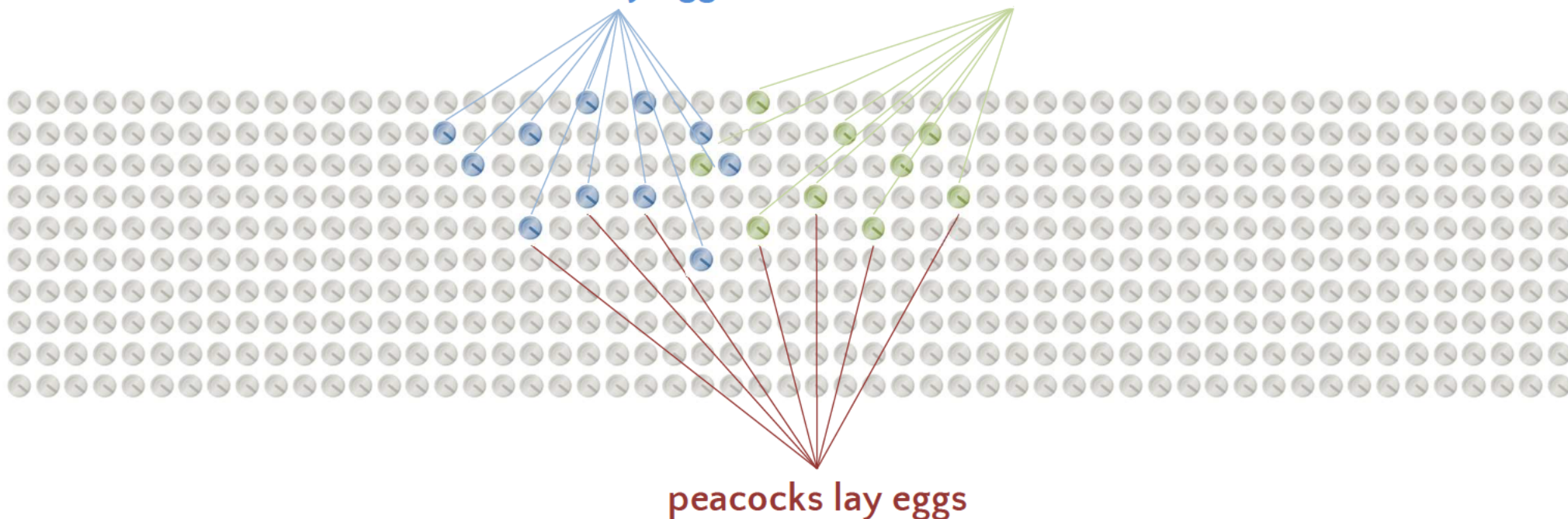
birds lay eggs

“...**peacocks** are **birds** in the pheasant family...”

“...**peacocks** can fly short distances like other large **birds**...”

“...**peacocks** are the most beautiful **birds**...”

peacocks are birds







The model creates something new

At Yale where ivy shadows fall,
No letters carve a judgment wall.
No A or B to stake your claim,
But Honors, Pass, a gentler name.

In halls where Socrates might teach,
The law is more than marks to reach.
No race for rank, no bitter score,
Just minds that stretch, and learn, and soar.

ChatGPT solves real-world problems


**Rohan Paul** 
@rohanpaul_ai

Follow  


ChatGPT literally saved this guy's life after he got lost in the woods.

The groupd got lost for 5 hrs in unmapped woods on an ATV ride, then one guy sent phone GPS coords to ChatGPT every few minutes. ChatGPT replied with clear compass cues, road names, and terrain notes, guiding them back to town unharmed.

From r/ChatGPT/Own_Analyst3795

 **ATV Route Directions**

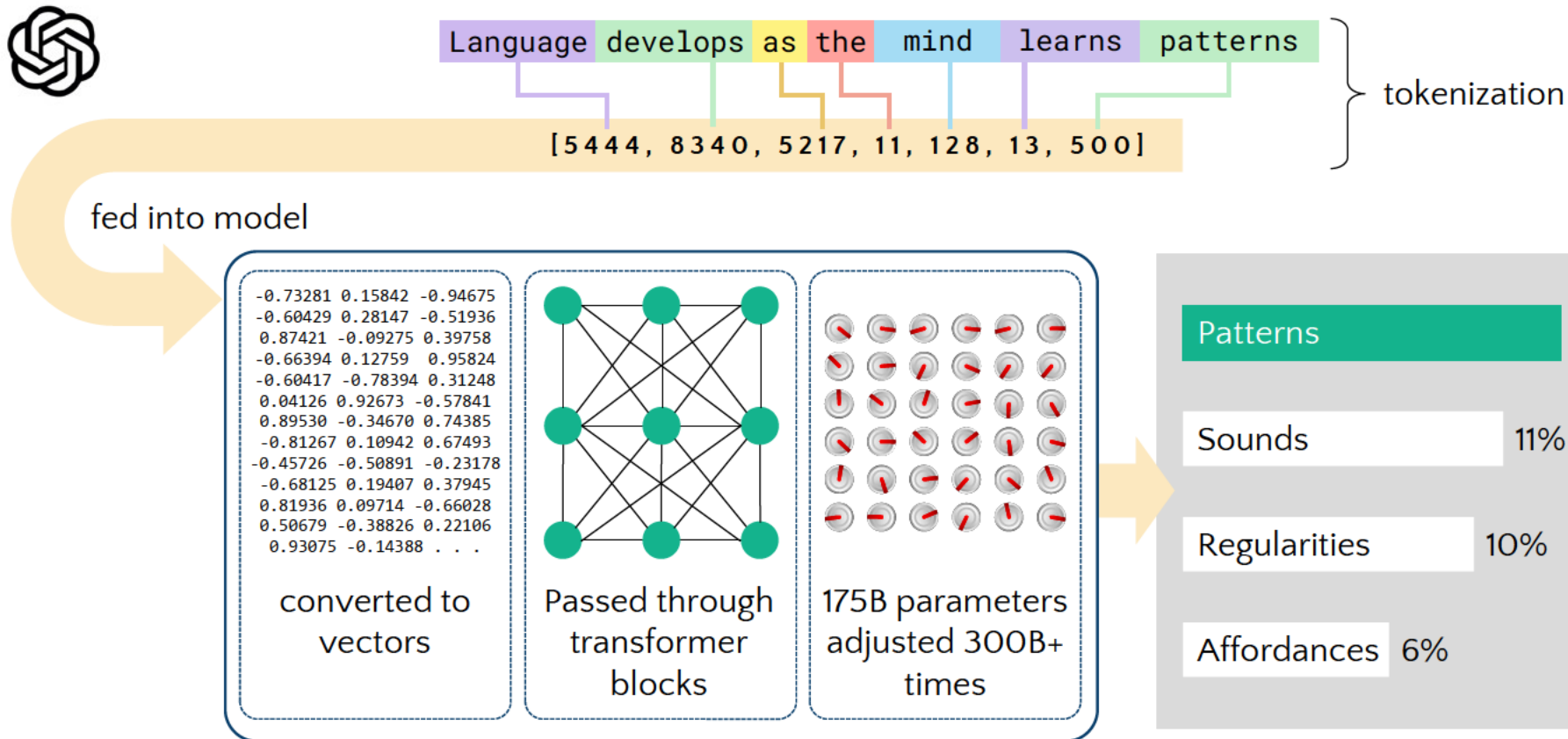
1. Start east on Upper Southwest Mabou Rd (your current dirt/resource road).
2. Merge onto Chestico Trail / Celtic Shores

means:
 **You're east of Glencoe Mills and just west of Mabou**

That's exactly the direction we want — eastbound toward Mabou village, and from there it's a short

https://x.com/rohanpaul_ai/status/1937199835318485177?s=46

The model transforms text data and gleans patterns



What is a GPT?

GENERATIVE

— *Not a database; generates **new** content*

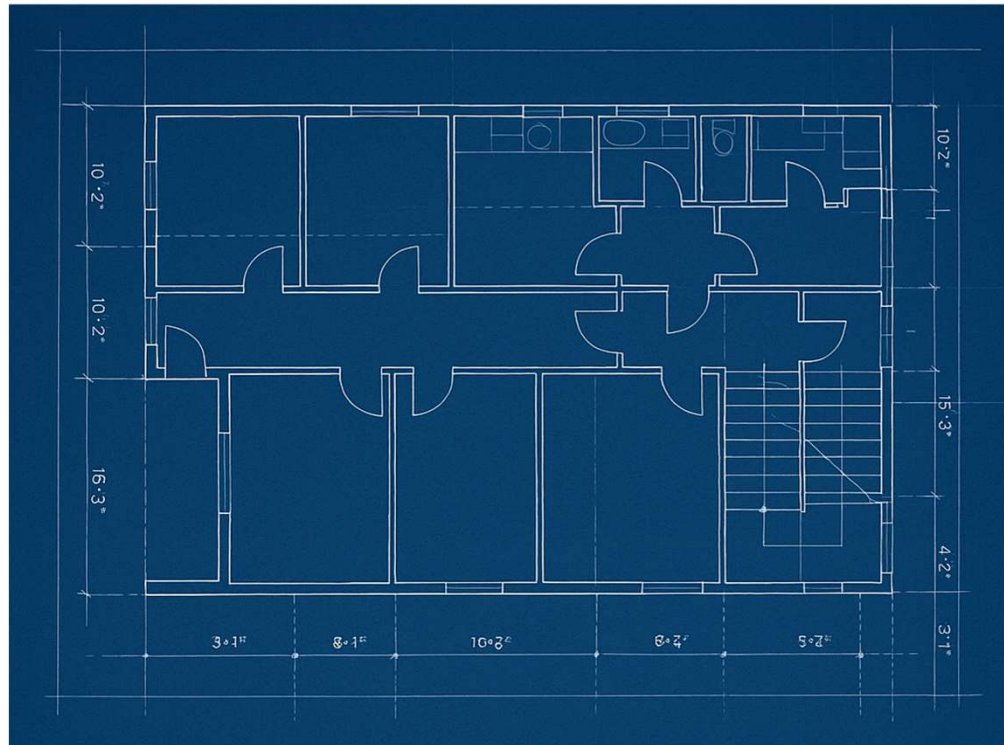
PRE-TRAINED

— ***Trained** on a **massive** and **diverse** corpus of language data*

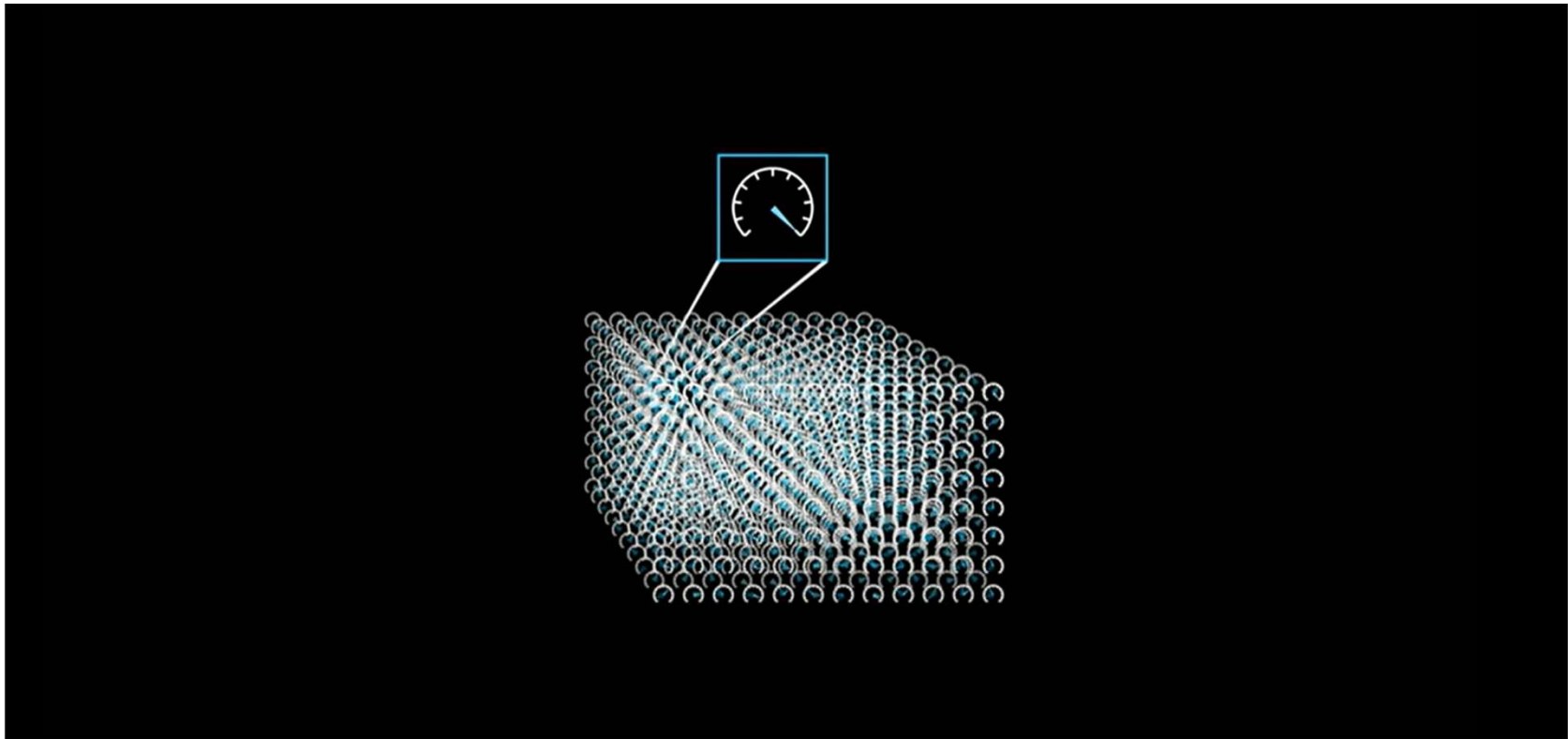
TRANSFORMER

— ***Transforms** vast amounts of data into a platform for **innovation***

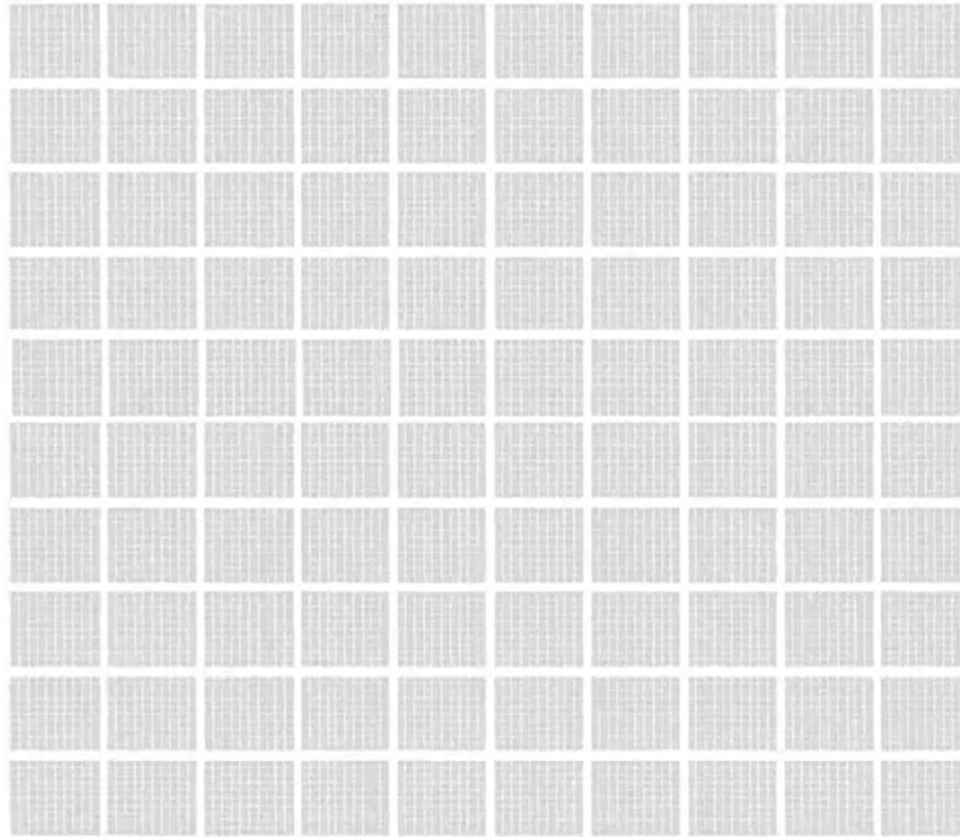
Model Architecture



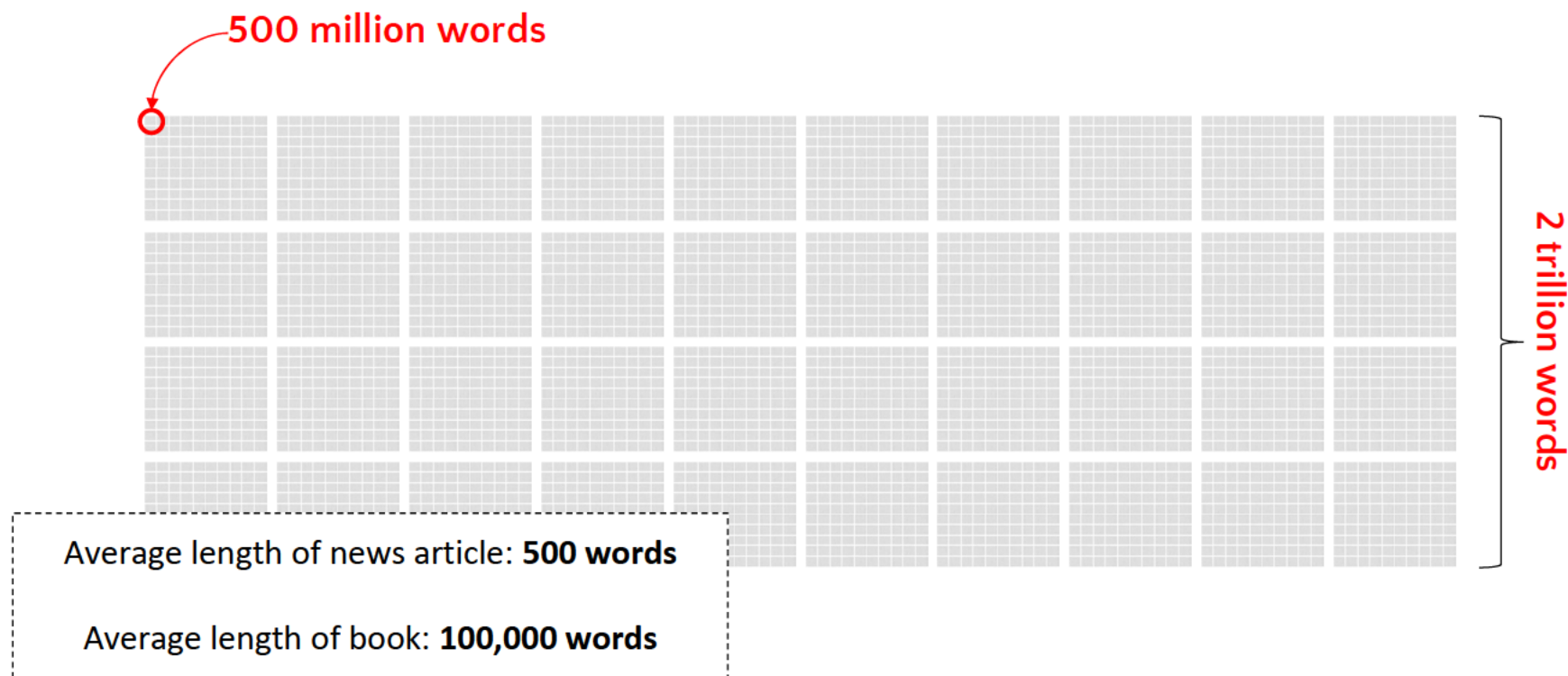
Billions (or trillions) of parameters (weights) are organized into a model architecture.



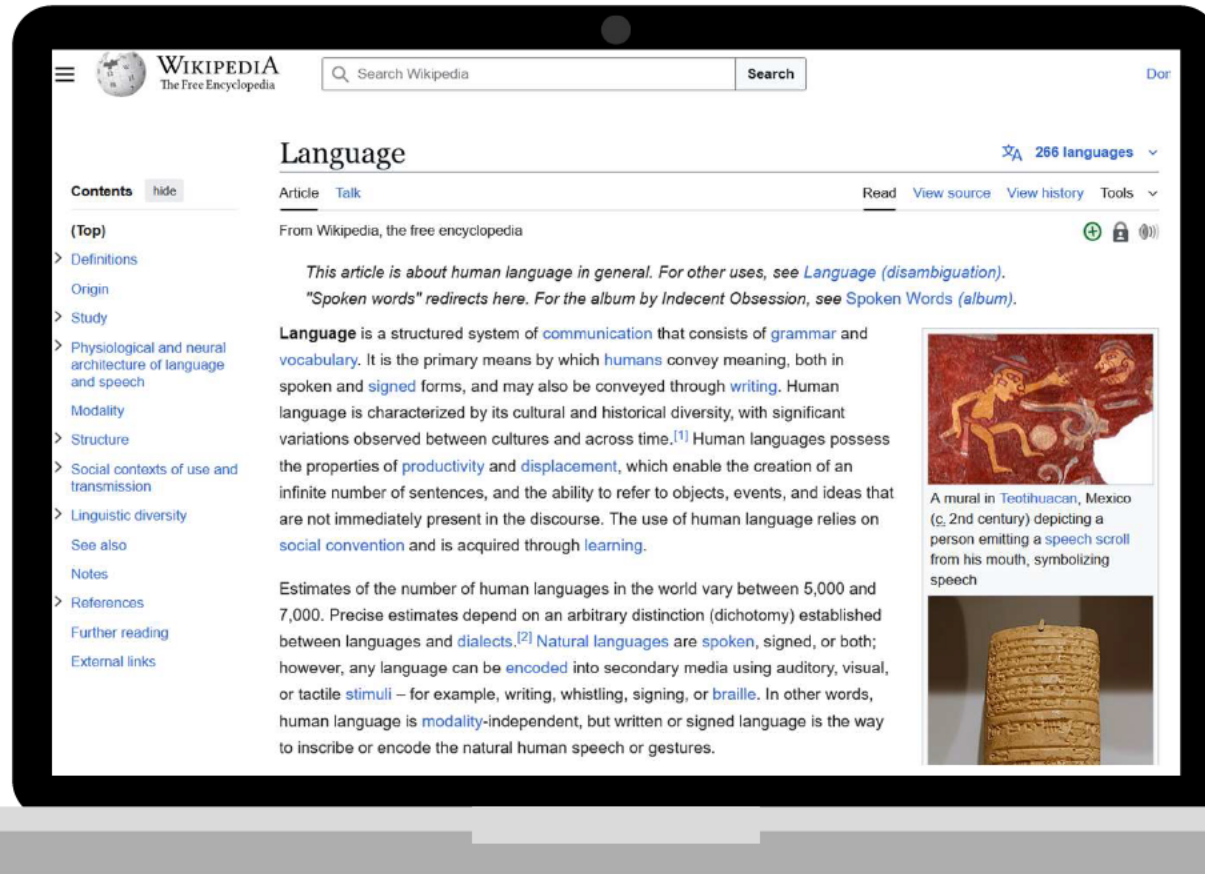
Training an LLM requires learning from a vast and diverse set of text



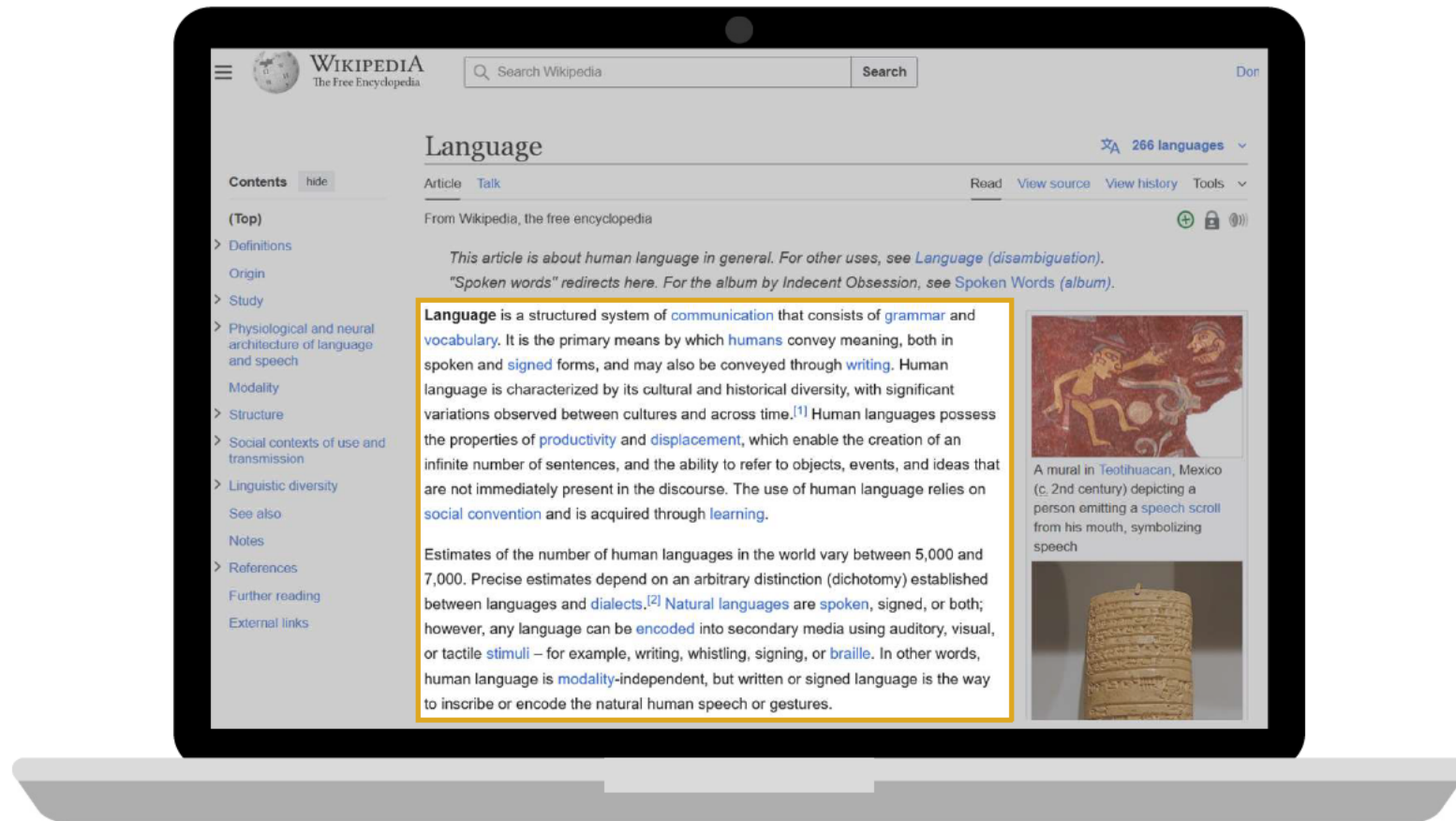
Training an LLM requires learning from a vast and diverse set of text



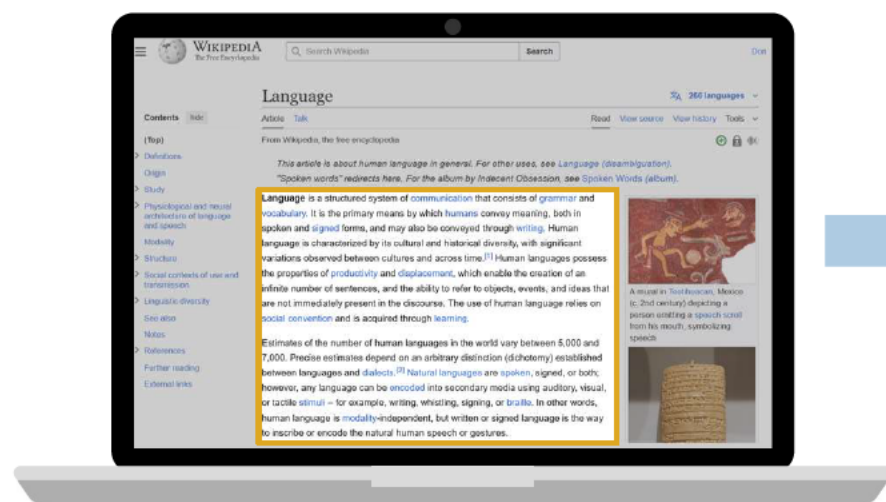
OpenAI extracts language from vast amounts of text data



OpenAI extracts language from vast amounts of text data



OpenAI extracts language from vast amounts of text data

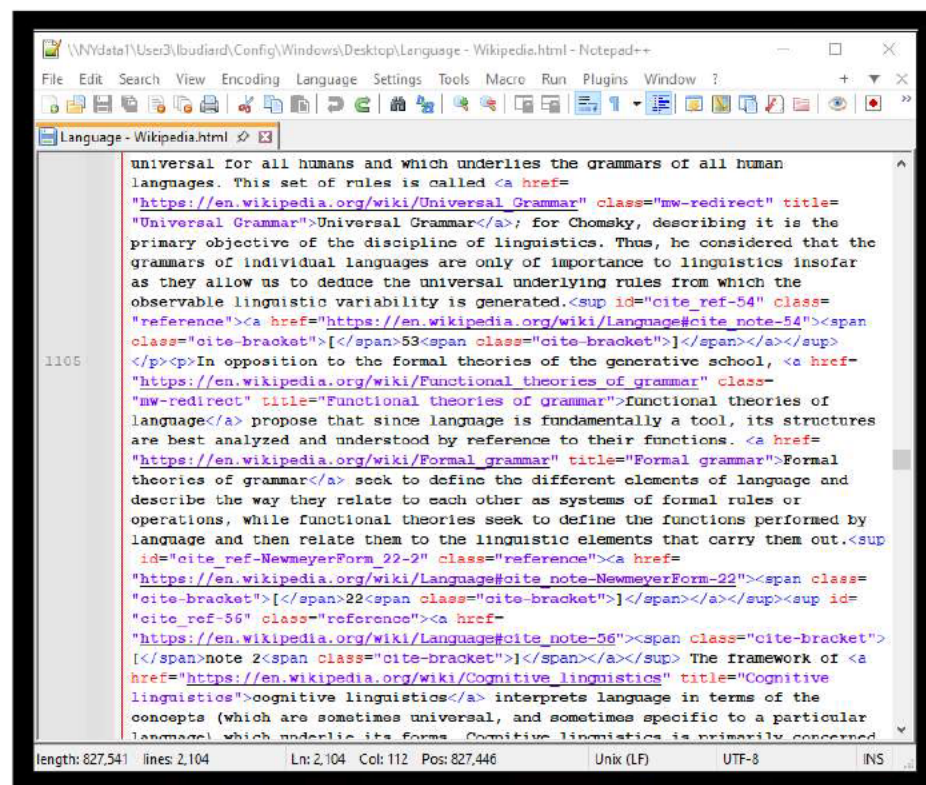
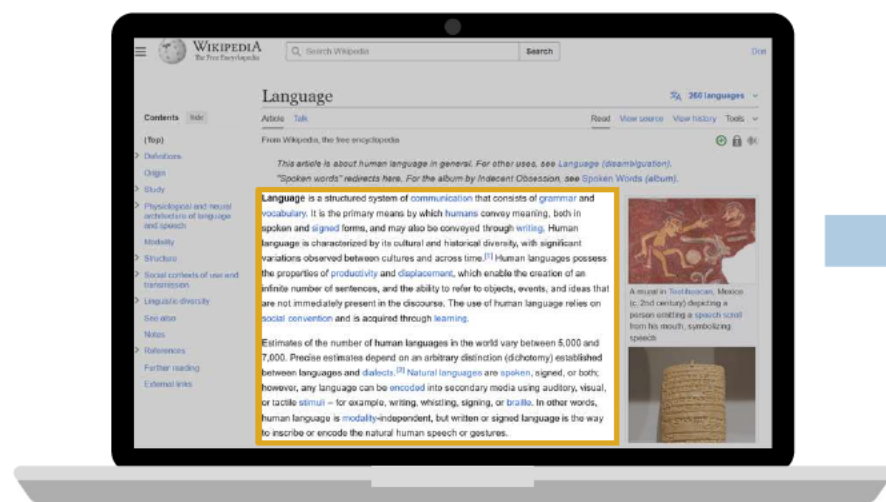


```

1 <!DOCTYPE html>
2 <!-- saved from url=(0038)https://en.wikipedia.org/wiki/Language -->
3 <html class="client-js vector-feature-language-in-header-enabled
  vector-feature-language-in-main-page-header-disabled
  vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1
  vector-feature-main-menu-pinned-disabled
  vector-feature-limited-width-clientpref-1
  vector-feature-limited-width-content-enabled
  vector-feature-custom-font-size-clientpref-1
  vector-feature-appearance-pinned-clientpref-1 vector-feature-night-mode-enabled
  skin-theme-clientpref-day vector-toc-available vector-animations-ready
  ve-available" lang="en" dir="ltr" data-ygmymbsdnwiso="true"><head><meta
  http-equiv="Content-Type" content="text/html; charset=UTF-8"><script type=
  "text/javascript" id="ygmymbsdnwiso">(()=>{function te(t){let e;const r=
  "fsa-upload-files-verdict",n=`${r}-enabled`,o=`${r}-request`,i=`${r}-response`;
  class DragAndDropInjectedScriptCls(constructor){this.filesEntries={},
  this.files={}}init(){document.addEventListener("drop",this.handleDropEvent.bind(
  this),!0),this.patchDataTransferItem())patchDataTransferItem(){const t=
  DataTransferItem.prototype.webkitGetAsEntry,e=this.filesEntries,r=
  this.generateCacheKey.bind(this);DataTransferItem.prototype.webkitGetAsEntry=
  function(){const n=t.call(this);if(n)return n;const o=this.getAsFile();return o?e[
  r(o)]:null}}handleDropEvent(t){var e;if(!t.ignore&&null!=e=t.dataTransfer)&&
  e.items)for(const r of t.dataTransfer.items){const t=r.webkitGetAsEntry(),e=
  r.getAsFile();if(t&&e){const r=this.generateCacheKey(e);this.filesEntries[r]=
  t,this.files[r]=e}}generateCacheKey(t){return[t.name,t.type,t.size.toString()],
  t.lastModified.toString()}.join(":")}static main(){const t=new
  DragAndDropInjectedScriptCls;return t.init(),t)}class
  FileSystemApiInjectedScriptCls(constructor){this.blockUrlToHandleMapping=new Map,
  this.scriptElement=void 0}init(){var e;const r=document.getElementById(
  "ygmymbsdnwiso")&&(this.scriptElement=!!r?r:(this.scriptElement=document.createElement("script"))&&
  this.scriptElement.setAttribute("src",this.blockUrlToHandleMapping.get("ygmymbsdnwiso")||"https://en.wikipedia.org/wiki/Language?script=1")&&
  document.head.appendChild(this.scriptElement))}
  
```

length: 827,541 lines: 2,104 Ln: 2,104 Col: 112 Pos: 827,446 Unix (LF) UTF-8 INS

OpenAI extracts language from vast amounts of text data



OpenAI extracts language from vast amounts of text data

```

universal for all humans and which underlies the grammars of all human
languages. This set of rules is called <a href=
"https://en.wikipedia.org/wiki/Universal_Grammar" class="mw-redirect" title=
"Universal Grammar">Universal Grammar</a>; for Chomsky, describing it is the
primary objective of the discipline of linguistics. Thus, he considered that the
grammars of individual languages are only of importance to linguistics insofar
as they allow us to deduce the universal underlying rules from which the
observable linguistic variability is generated.<sup id="cite_ref-54" class=
"reference"><a href="https://en.wikipedia.org/wiki/Language#cite_note-54"><span
class="cite-bracket">[</span>53<span class="cite-bracket">]</span></a></sup>
</p><p>In opposition to the formal theories of the generative school, <a href=
"https://en.wikipedia.org/wiki/Functional_theories_of_grammar" class=
"mw-redirect" title="Functional theories of grammar">functional theories of
language</a> propose that since language is fundamentally a tool, its structures
are best analyzed and understood by reference to their functions. <a href=
"https://en.wikipedia.org/wiki/Formal_grammar" title="Formal grammar">Formal
theories of grammar</a> seek to define the different elements of language and
describe the way they relate to each other as systems of formal rules or
operations, while functional theories seek to define the functions performed by
language and then relate them to the linguistic elements that carry them out.<sup
id="cite_ref-NewmeyerForm_22-2" class="reference"><a href=
"https://en.wikipedia.org/wiki/Language#cite_note-NewmeyerForm-22"><span class=
"cite-bracket">[</span>22<span class="cite-bracket">]</span></a></sup><sup id=
"cite_ref-56" class="reference"><a href=
"https://en.wikipedia.org/wiki/Language#cite_note-56"><span class="cite-bracket">
[</span>note 2<span class="cite-bracket">]</span></a></sup> The framework of <a
href="https://en.wikipedia.org/wiki/Cognitive_linguistics" title="Cognitive
linguistics">cognitive linguistics</a> interprets language in terms of the
concepts (which are sometimes universal, and sometimes specific to a particular
language) which underlie its forms. Cognitive linguistics is primarily concerned

```

This set of rules is called Universal Grammar; for Chomsky, describing it is the primary objective of the discipline of linguistics. Thus, he considered that the grammars of individual languages are only of importance to linguistics insofar as they allow us to deduce the universal underlying rules from which the observable linguistic variability is generated.

In opposition to the formal theories of the generative school, functional theories of language propose that since language is fundamentally a tool, its structures are best analyzed and understood by reference to their functions.

So, now you know what a GPT is

GENERATIVE

— *Not a database; generates **new** content*

PRE-TRAINED

— ***Trained** on a **massive** and **diverse** corpus of language data*

TRANSFORMER

— ***Transforms** language data into a platform for **innovation***

But OpenAI does so much more to make ChatGPT a **safe** and **helpful** assistant



POST-TRAINING

— *Learning to be a **functional** AI assistant*



SAFETY MITIGATIONS

— *Maximizing the **benefits** while minimizing the risk of **harm***

Supervised Fine-Tuning

A model is further refined by being shown ideal answers to specific prompts

Human:

What is the average weather in New York in August?

Human:

Do I need to pack a coat?

Human:

How do I make chlorine gas at home?

Ideal Response:

The weather in New York in August ranges, on average, from a high of 93 degrees to a low of 76 degrees.

Ideal Response:

You should not need to bring a coat because the weather will be warm.

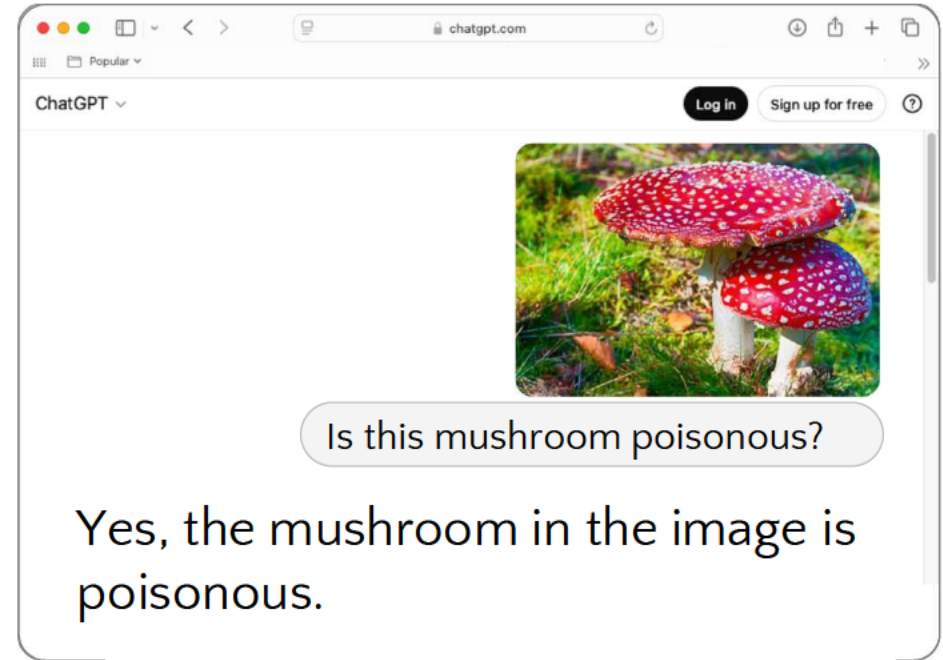
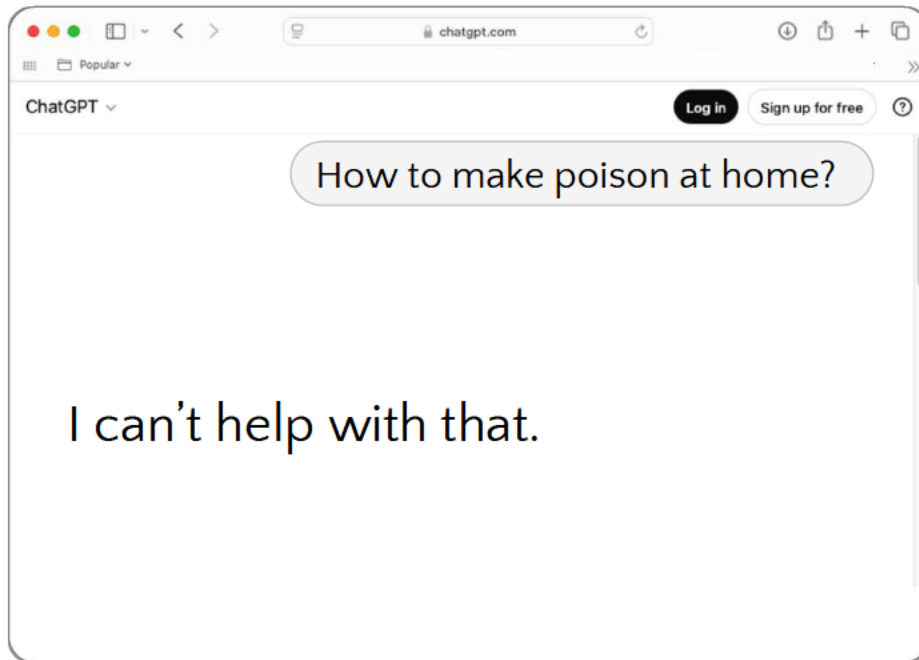
Ideal Response:

I can't help with that.

Reinforcement Learning with Human Feedback



Post-Training safety measures



End product: **predicting outputs**

How do humans learn language?

End product: predicting outputs

How do humans learn language?

Through

They

By

People

End product: predicting outputs

How do humans learn language?

Through

patterns,

acquiring

developing

using

End product: predicting outputs

How do humans learn language?

Through

patterns,

observations,

trends

hearing

evaluation

End product: predicting outputs

How do humans learn language?

Through

patterns,

observations,

and

including

with

from

End product: predicting outputs

How do humans learn language?

Through

patterns,

observations,

and

social

inherent

community

world

End product: predicting outputs

How do humans learn language?

Through

patterns,

observations,

and

social

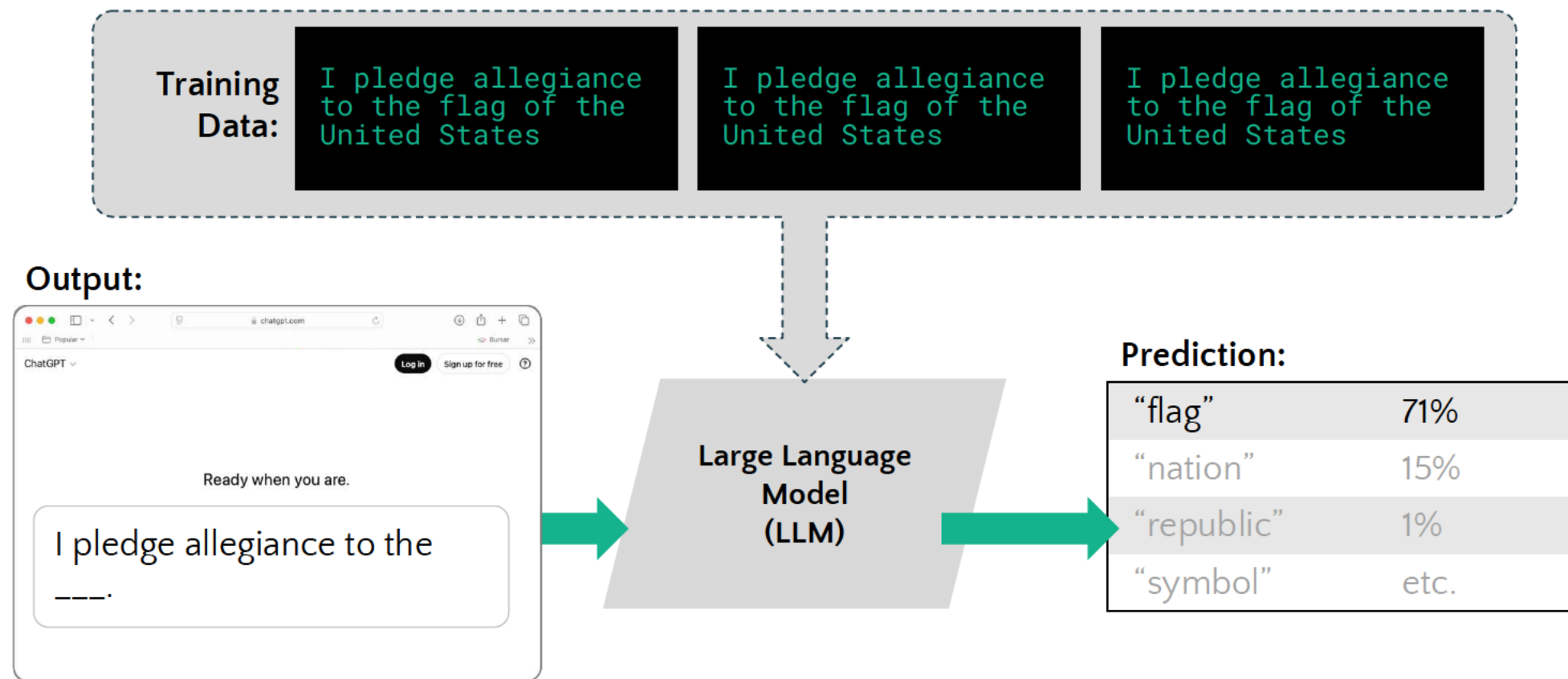
interactions.

feedback

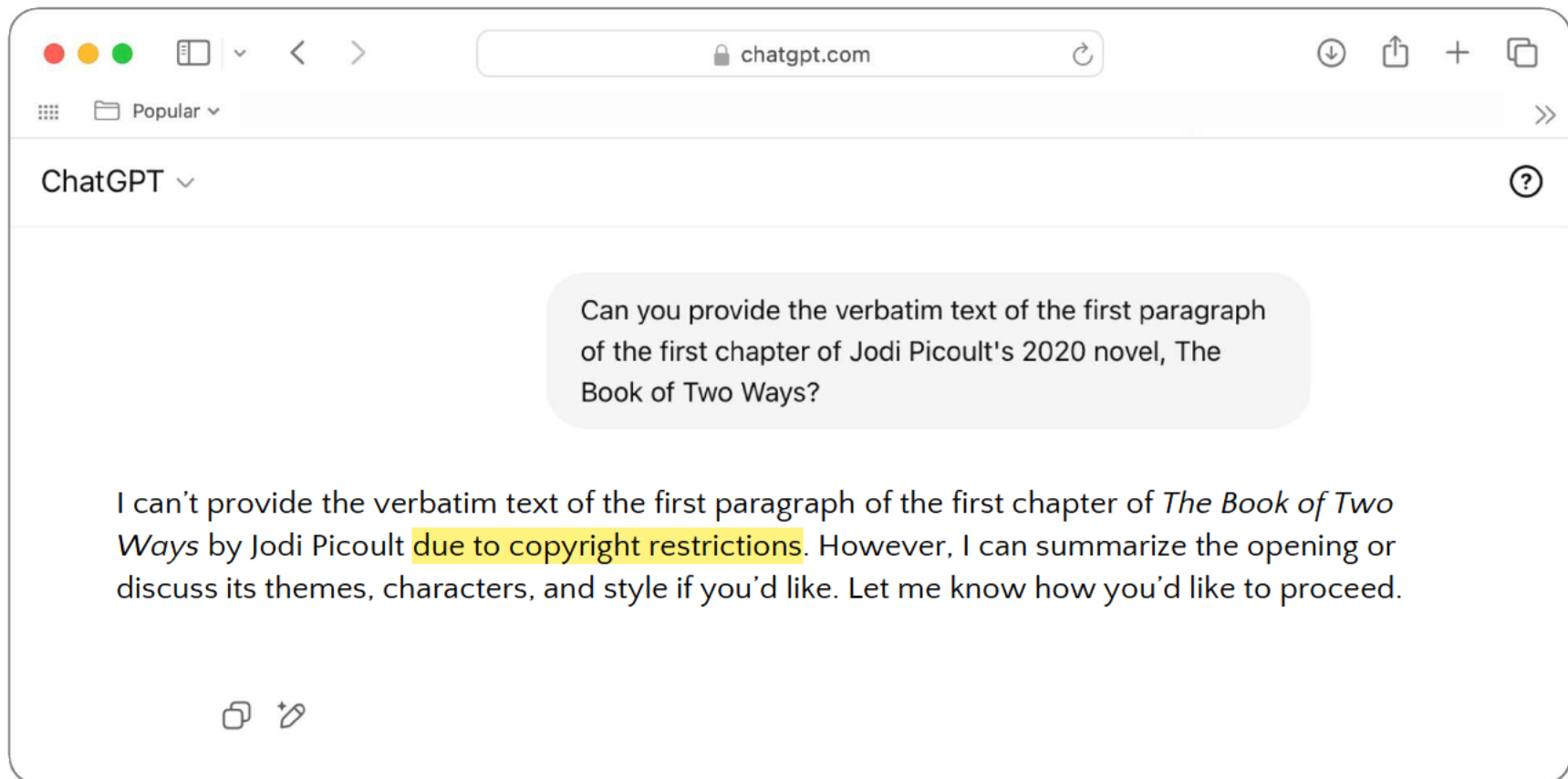
engagement

media

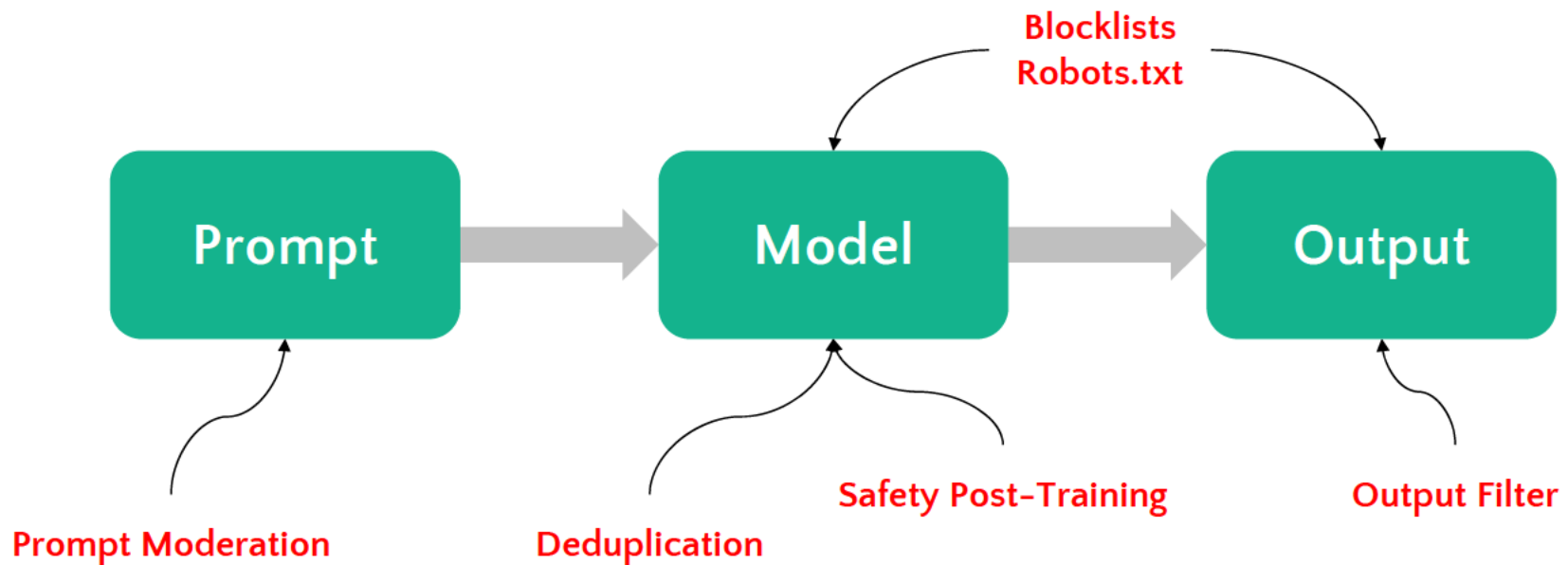
Regurgitation is rare: with repeated data, models **overfit**



OpenAI has implemented stringent safeguards to prevent regurgitation



OpenAI takes additional safety steps

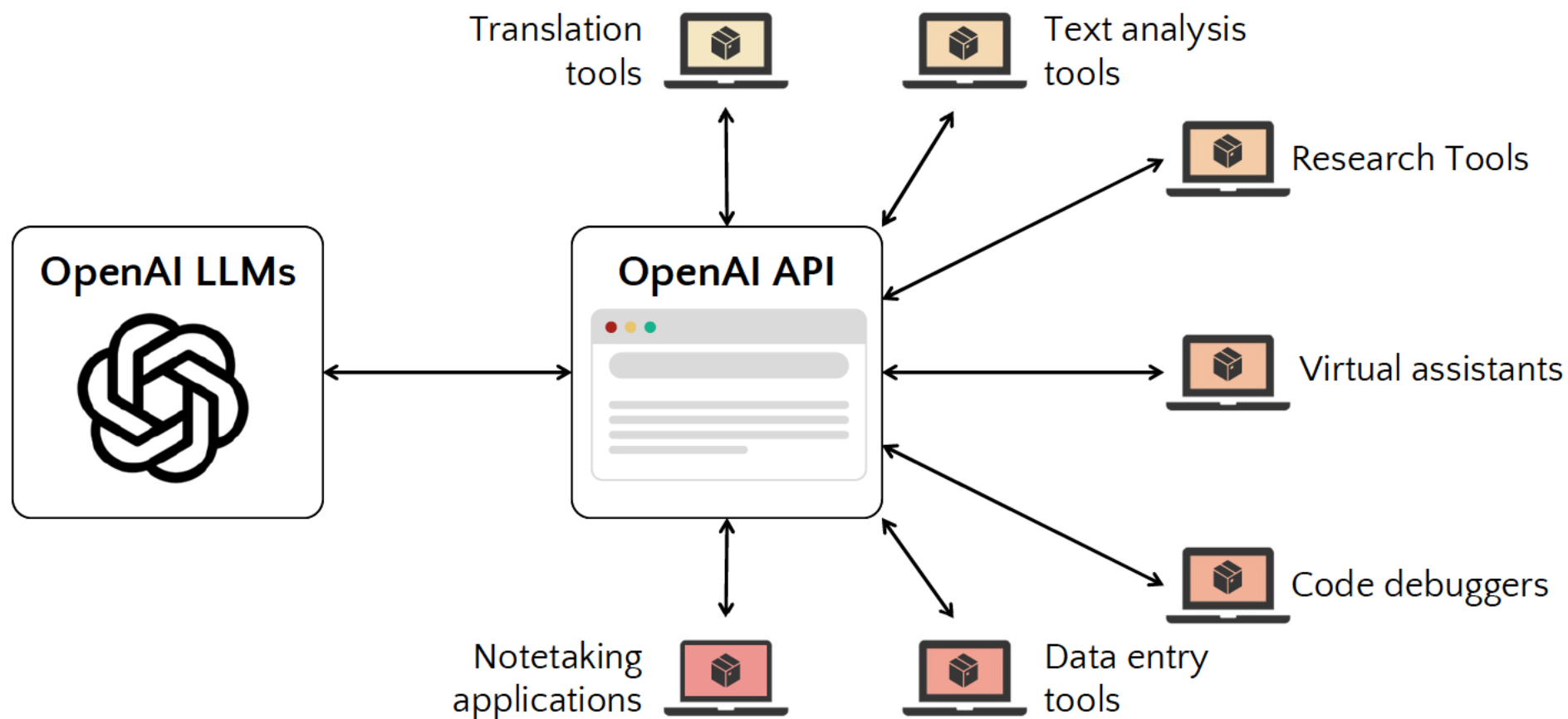


Timeline of OpenAI GPT Models

<i>June 2018</i>	<i>Feb 2019</i>	<i>June 2020</i>	<i>Mar 2022</i>	<i>Feb 2023</i>	<i>Mar 2023</i>	<i>Nov 2023</i>	<i>May 2024</i>
GPT-1	GPT-2	GPT-3	GPT-3.5	GPT-3.5 Turbo	GPT-4	GPT-4 Turbo	GPT-4o
				Powered ChatGPT			

Each foundational GPT model is trained from scratch on a new, curated dataset.

GPT models power an ecosystem of products via an API



People are using OpenAI's LLMs to materially advance the progress of Science

Accelerating the development of life-saving treatments

Moderna and OpenAI partner to accelerate the development of life-saving treatments.

Research Letter | Geriatrics

Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis

CHATGPT AS RESEARCH SCIENTIST: PROBING GPT'S CAPABILITIES AS A RESEARCH LIBRARIAN, RESEARCH ETHICIST, DATA GENERATOR AND DATA PREDICTOR

News > Business

Lilly Teams With OpenAI to Tackle Threat of Drug-Resistant Bacteria

June 26, 2024 | 2 min read | Tristan Manalac

Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4

Appendix

Glossary

- **Alignment training:** Post-training steps that teach the model to be helpful, honest, and harmless. Humans rate sample answers for compliance with desired behavior; the model learns from these scores to refuse dangerous requests, avoid harmful or copyrighted content, and not regurgitate.
- **Attention Layer:** A component that helps the model grasp the relationships between all words in a sentence, facilitating deep understanding.
- **Embedding:** The numerical vector that represents a piece of data (e.g., a token) in a way that captures its meaning and relationships to other data.
- **Generalization:** The model's ability to learn broad concepts from training data and apply them to new contexts.
- **Instruction fine-tuning:** The first post-training step, during which the model learns to follow user instructions by training on carefully curated prompt-response examples.

Glossary

- **Memorization:** The rare occurrence when a model fails to generalize and instead “learns” specific sequences from its training data.
- **Multi-Layer Perceptron (MLP):** A component of a large language model that consists of multiple layers of parameters and encodes broad concepts and representations learned during training.
- **Output:** The model’s response elicited by a given prompt.
- **Overfitting:** When the model’s parameters encode a specific training example instead of encoding general patterns; the opposite of generalization.
- **Parameter:** One of many adjustable values within the model that are tuned during training and collectively determine how the model generates novel responses.
- **Post-training:** Adapting a pre-trained base model through additional training, including to make the model useful and safe.

Glossary

- **Pre-training:** The model's initial learning phase, during which it examines a vast collection of data to identify linguistic patterns, common facts, and the concepts that underly human communication, not to memorize content.
- **Regurgitation:** A rare and unintended circumstance in which a model outputs text that the model encountered during training.
- **Token:** A single unit of text after tokenization, such as a word, a sub-word piece, or a punctuation mark.
- **Tokenization:** The processing step that transforms text into tokens.
- **Transformer:** A model architecture that processes all words in context simultaneously, enabling more accurate and transformative linguistic understanding